Submitted to the 17th Annual International Conference on Intelligent Systems for Molecular Biology
& 8th European Conference on Computational Biology (ISMB/ECCB 2009).
draft material -- please do not cite

Category: Databases and Ontologies

Alignment of the UMLS Semantic Network with BioTop. Methodology and Assessment.

Erik M. van Mulligen[1], László van den Hoek[1], Olivier Bodenreider[2], Elena Beisswanger[3] , Stefan Schulz[4],

[1]Erasmus MC, Rotterdam, The Netherlands

[2]National Library of Medicine, Bethesda, MD, USA

[3]Jena University Language and Information Engineering (JULIE) Lab, Germany

[4]Institute of Medical Biometry and Medical Informatics, University Medical Center Freiburg, Stefan-Meier-Str. 26, 79104 Freiburg, Germany

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

Motivation:

For many years, the UMLS Semantic Network (SN) has been used as an upper-level semantic framework for the categorization of terminological resources in biomedicine. BioTop has recently been developed as an upper-level ontology for the biomedical domain. In contrast to the SN, it is founded upon strict ontological principles, using OWL-DL a formal representation language, which has become popular through the Semantic Web. In order to make logic-based reasoning available for the resources annotated or classified by the SN, a mapping ontology was developed aligning the SN with BioTop.

Methods:

The theoretical foundations and the practical realization of the alignment are described, with a focus on the design decisions taken, the problems encountered, and the adaptations of BioTop that became necessary. For evaluation, UMLS concept pairs obtained from MEDLINE abstract sentences by a named entity recognition system were tested for possible semantic relationships. Furthermore, all semantic type combinations which occur in the UMLS Metathesaurus were checked for consistency.

Results:

The effort-intensive alignment process required major design changes and enhancements of BioTop and brought up several design errors that could be fixed. A comparison between a human curator and the ontology yielded only a low agreement. Ontology reasoning was also used to successfully identify 89 inconsistent semantic type combinations.

Availability:

BioTop, the OWL-DL representation of the UMLS SN, and the mapping ontology are available at http://purl.org/biotop/

Contact: stschulz@uni-freiburg.de

1 INTRODUCTION

As high-throughput experimental methods and advanced information technology have impressively increased the amount of data, the resulting information congestion has well-known consequences such as fragmentation of data and knowledge and duplication of research efforts. Especially the modal, spatial, and semantic fragmentation of *protein information* is a major impeding factor for biological research [Stevens2000].

Factual information about proteins, genes, diseases and other relevant biomedical entities are increasingly available in structured databases but its dissemination by unstructured, i.e. research articles, still prevails. It is estimated that as much as 80% of new scientific facts is communicated only in their original journal publication [Jelier2005], the authors relying on a limited group of curators to manually extract, annotate, and transfer these facts into the appropriate databases.

Although the pooling of such facts in databases like UniProt [Mulder2008], offers great advantages over the traditional publication process, it would be of great benefit to concentrate all this information in a structured manner in one centralized repository: ongoing research information, peer-reviewed articles, external, authoritative knowledge bases, together with formalizations of the basic kinds of entities and their interrelations in so-called ontologies. This is one of the goals that the WikiProteins [Mons2008] project aims to accomplish. WikiProteins proposes to open the curation process to the community at large. This will shift the burden of adding and improving the information away from the commonly understaffed curators and towards the scientific community, which is expected to result in a more complete, up-to-date and accurate collection of information than can be achieved with current means.

WikiProtein uses natural language processing techniques for identifying named entities in text and annotating them to reference terminologies such as the Unified Medical Language System

(UMLS). Corrections can then be entered by users directly into the WikiProtein database, improving named entity recognition.

The structured information in WikiProtein can be exploited by automatic reasoning services for tasks including hypothesis generation and knowledge discovery. Reasoning services, however, require sound ontologies and may produce suboptimal results when based on traditional terminological systems. For this reason, we set out to examine the extent to which using a formal domain ontology (BioTop) instead of the legacy UMLS Semantic Network changes the interpretation that can be made of potential assertions extracted from WikiProtein. More precisely, we created a mapping between the UMLS Semantic Network [McCray2003] and BioTop [Schulz2006], and assessed through this mapping how each resource contributes to the interpretation of the relation between pairs of co-occurring concepts.

The paper is organized as follows: After an overview of basic concepts like terminology and ontology underlying our work (section 2) we describe the resources used, the mapping approach, and the evaluation methodology (section 3). We present then our results and discuss them in the context of related work (sections 4 and 5).

2 BACKGROUND

We here introduce the basic concepts underlying our work, viz. terminology, ontology, and description logics.

2.1 Terminology

Both text mining and manual annotation require some kind of semantic standard. Originally this issue was supposed to be addressed by controlled vocabularies and terminology systems [ISO2000, DeKeizer2000a, DeKeizer2000b,], a heterogeneous group of mostly language-oriented artifacts that relate the various senses or meanings of linguistic entities with each other, (e.g. by recording the synonymy between "Nephroblastoma" and "Wilms' Tumor"). Sets of (quasi-) synonymous terms are commonly referred to as 'concepts', and in many terminology systems concepts are furthermore related by informal semantic relationships often following vague natural language predicates (*narrower than*, *associated with*,…). Terminology systems are generally built to serve a well-defined purpose such as document retrieval, resource annotation, the recording of mortality and morbidity statistics, or billing. In the medical field, the largest terminological system is the Unified Medical Language System [UMLS2009,Bodenreider2004] in which synonymous terms from different source vocabularies are clustered into concepts, each of which is categorized using a system of semantic types [McCray1995]. Today, the UMLS comprises 1.9 million concepts and almost seven million terms from close to 150 sources.

2.2 Ontology

In reaction to the language- and purpose-oriented and informal approaches to representing a given domain, there has been a growing interest in using formal methods for precisely describing the invariant and language-independent properties of the entities in a domain. In biomedicine, the Gene Ontology (GO) [Ashburner2000] was the pioneer of moving from a purpose-oriented annotation vocabulary to a more principled resource. Similarly, collaborative initiatives have emerged such as the Open Biomedical Ontologies (OBO) Foundry [Smith2007], the continuing development of SNOMED CT [SNOMED2009], which is increasingly challenged and guided by ontological principles, as well as increasing mutual awareness between the Semantic Web and Life Sciences communities [Ruttenberg2007,Sagotsky2008].

The term "ontology" stems from analytical philosophy, concerned with the question of "what exists?" [Quine1948]. It became popular by information sciences, and in spite quite contradictory definitions [Kusnierczyk2006] it has increasingly been used to refer to domain representation of various kinds. In order to emphasize the use of a formal language in domain representations, we here subscribe to the concept of *formal ontologies* [Guarino1998], as theories that attempt to give precise representations of the types of entities in reality, of their properties and of the relations between them, using axioms and definitions that support algorithmic reasoning..

2.3 Upper level Ontologies

The purpose of upper domain ontologies is to define the foundational kinds and relations relevant to the entire domain. In Life Sciences, such classes include gene, protein, cell, tissue, nucleotide, population, organism, diagnostic procedure and biological function. Upper domain ontologies can either be used alone as a source of basic categories (e.g., for the coarse annotation of resources) or as a common reference for more specialized domain ontologies.

In contrast to domain-specific ontologies, such as the Gene ontology, upper ontologies propose to trade detail for scope by introducing general categories that are the same across all domains. Whether or not this is achievable and desirable has been subject of debate. Nevertheless, several upper-level ontologies have been developed and are being maintained such as BFO[1] [Smith2007a], DOLCE[2] [Gangemi2002,Masolo2003], SUMO[3] [Pease2008], or GFO[4] [Heller2004], but the development of application-oriented domain ontologies such as the OBO[5] ontologies has led to the proposal of a kind of intermediate-level ontologies, also called top-domain ontologies, such as the Simple Bio Upper Ontology [Rector2006], GFO-Bio [Hoehndorf2009], or BioTop [Schulz2006]. In contrast to these recent and more theoretical resources, the UMLS[6] Semantic Network (UMLS SN), developed fifteen years ago, can be regarded as the archetype of a biomedical domain upper ontology [McCray2003]. Moreover, the Semantic Network has already proved its usefulness in providing a consistent categorization of all concepts represented in the UMLS Metathesaurus.

From an upper-level ontology viewpoint, domain upper ontologies play the role of domain ontologies, but from a domain perspective they act as upper ontologies. For example, the placement of BioTop under BFO could be seen as a domain ontology placed under an upper ontology. Conversely, BioTop itself may also play

[1] Basic Formal Ontology

[2] Descriptive Ontology for Linguistic and Cognitive Engineering

[3] Suggested Upper Merged Ontology

[4] General Formal Ontology

[5] Open Biomedical Ontologies

[6] Unified Medical Language System

the role of an upper ontology when linked to the Cell Ontology or the Gene Ontology.

Different upper level ontologies not only use different formalisms for their representation, but also model the domain in slightly different ways. As a consequence, the constraints they impose on domain-specific ontologies affect the result of reasoning services based on these upper level ontologies.

2.4 Description Logics

Since the 1980's, the application of formal reasoning on ontology structures has led to various formalisms. Later on, the vision of the Semantic Web [BernersLee2001] has resulted in a significant standardization of representation languages, formats and reasoning engines.

One of the most noteworthy standards of the Semantic Web was the development of the web ontology language OWL [Horrocks2003], and especially its expressive but still computable subset, OWL DL. Description logics (DLs) constitute a family of decidable fragments of first-order logic which have a clean and intuitive syntax [Baader2007]. They come in various flavors, ranging from lightweight to highly expressive ones. The trade-off between expressivity of the logic and computability (and thus, scalability) of its reasoning has to be made in order to properly address the ontology application. Whereas overly inexpressive DL may lead to underspecifications that imply unintended models of the ontology, highly expensive reasoning makes it infeasible from practical viewpoints. OWL-DL constitutes a compromise between expressiveness and decidability and is supported by so-called classifiers like RACER, Fact++, Pellet [Haarslev2003, Tsarkov2006, Sirin2007].

Description logics are built around the notions of "class" and "relationship" and follow model-theoretic semantics. Classes such as *Heart* are interpreted as sets of all instances belonging to that class, i.e., here: all particular hearts in the domain. Relationships then are sets of pairs of class instances like *has_part,* which extends to all pairs of objects in the domain that are related in terms of parts and wholes. So are all pairs of heart instances with their respective mitral valve instances in the extension of *has_part*. We will illustrate DL syntax and semantics through a set of increasingly complex examples, starting with the class *Liver*, which, in our domain extends to all individual livers of all organisms. Analogously, the class *BodilyOrgan* then extends to all individual bodily organs. When those two statements are put together, we can introduce the key concept of taxonomic subsumption: The class *BodilyOrgan* forms a superclass of the class *Liver*, i.e., the former subsumes the latter if and only if all particular livers are also instances of the class *BodilyOrgan*. In DL notation, this taxonomic subsumption is expressed by the $\sqsubseteq$ operator, e.g., *Liver* $\sqsubseteq$ *BodilyOrgan*, and is also known as subtype, subclass, or *is-a* relationship. It is important to stress that this kind of relationship always relates two classes only. In contradistinction to this, the instantiation relationship always relates an individual entity to some class, e.g., the particular (liver of the first author of this paper with the class *Liver*.

Such simple class statements can then be combined by different operators and quantifiers, e.g. the $\sqcap$ ("and") operator and the existential quantifier $\exists$ ("exists"). For example, *InflammatoryDisease* $\sqcap$ $\exists$*hasLocation.Liver* denotes all instances that belong to the class *InflammatoryDisease* and are further related through the relationship *hasLocation* to some instance of the class *Liver*. This example actually gives both necessary and sufficient conditions in order to fully define the class *Hepatitis*:

Hepatitis $\equiv$ *InflammatoryDisease* $\sqcap$ $\exists$*hasLocation.Liver*

The equivalence operator $\equiv$ indicates that every instance of hepatitis is necessarily an inflammatory disease that is located in some liver. But through the equivalence operator, one can go in the other direction as well and say that any inflammatory disease that is located in some liver can be classified as hepatitis. In practice, the term on the left and the expression on the right are equivalent.

3 MATERIALS AND METHODS

3.1 UMLS Semantic Network (SN)

The provision of an overarching conceptual umbrella over the biomedical domain was the rationale for the development of the UMLS SN [McCray2003]. A tree of 135 semantic types forms the backbone of the SN. It is partitioned into the branches "entity" and "event", in which nodes are linked by subclass relations. Additionally, the SN contains a hierarchy of 53 associative relationships (e.g., *location_of*, *treats*). These relationships are used to form 612 assertions (e.g., *Tissue, location_of, Diagnostic Procedure*), from which 6,252 additional assertions can be inferred. For each semantic relationship, domain and range are specified in terms of one or more semantic types. Each concept from the UMLS Metathesaurus is categorized by at least one semantic type from the SN.

3.2 BioTop

BioTop [Schulz2006] originated as a redesign and enrichment of the GENIA ontology. Like the UMLS SN, its backbone is constituted by a taxonomic tree, consisting of 283 classes. Its relation hierarchy is populated by 48 relations with domain and range constraints. The main difference to the UMLS SN is given by its use of OWL DL (see section 2.3). It contains more than 700 logical axioms, among which there are subclass , disjointness, and equivalence axioms. The latter (67) enable the computation of additional taxonomic links using DL reasoners. BioTop exhibits links to the upper level ontology BFO as well as to the OBO relation ontology [Smith2005]. Furthermore, it provides mappings to OBO Fourndry ontologies (e.g. Gene Ontology, Cell Ontology, FMA, ChEBI). A map to the DOLCE ontology [Gangemi2002] is currently under way.

3.3 WikiProteins

A major part of the data currently available in WikiProteins was acquired from the UMLS Metathesaurus. The original choice as an upper domain ontology for WikiProteins was therefore the UMLS Semantic Network, since each Metathesaurus concept is associated with one or more semantic types.

However, the UMLS Semantic Network suffers from some well-known shortcomings (ambiguous or vague class descriptions, relatively low granularity, arbitrary divisions) [Kumar2004]. For this reason, we wanted to assess these limitations by making them explicit in an OWL-DL representation and to explore alternative upper domain ontologies (e.g., BioTop).

3.4 Mapping

Our main objective of bridging between the UMLS SN and BioTop was to capitalize on the annotations by the former, especially in terms of the UMLS Metathesaurus on the one hand, and to benefit from the ontologically sound and computationally more sophisticated architecture of the latter. The aim was to represent the totality of the SN knowledge using BioTop, encompassing the SN types, the hierarchies, as well as the semantic relations with their domain and range restrictions. In order to meet this requirement, an analysis of the UMLS SN semantics in the light of description logics and its transformation into the formalism used by BioTop has to be performed. Technically, the plan was to use a central mapping file, which imported both UMLS SN and BioTop, and served as a store for class and relation equivalences and restrictions. As an agreement between WikiProteins and BioTop curators, the coverage of BioTop should be adjusted to STs wherever justified.

3.5 Assessment Methodology

- Formative Evaluation of BioTop: We used the logic-driven knowledge reengineering described by [Schulz2001], which consists in an iterative approach. Each major ontology redesign (including mapping) step is checked by a description logics reasoner, the results of which are then analyzed and corrected under two perspectives: Firstly, the classes tagged as "inconsistent" are identified, the causes are investigated and repaired. Secondly, as the ontology has reached a consistent state, the logical entailments are analyzed for adequacy. Whenever inadequate entailments are encountered, the causes are investigated and fixed.
- Named Entity Co-occurrence: Named entity recognition (NER) is a widely used text mining technique [Park2006] software. A well-known problem in NER is when the word of phrase to be recognized is ambiguous, i.e. it denotes different things. The implementation of the UMLS SN in BioTop offers the possibility to check ambiguous named entities for whether the competing referent concepts "fit together" according to the SN relations allowed for UMLS semantic types. In the context of WikiProteins we obtained ~100 Million unique pairs from ~15 Million PubMed abstracts that had been mined with state-of-art named entity recognizer Peregrine [Schuemie2007] to recognize UMLS concepts and Uniprot identifiers referred to within the same sentence. We here consider only the UMLS concept pairs. The task was to manually assessed a sample of ~300 UMLS concept pairs. The curator assessed the plausibility of the linkage between the two concepts in the sentence context. Each co-occurring pair was first checked against the SRSTRE1 table from the semantic network and alternatively against the mapping ontology, based on the OWL-DL implementation of the BioTop/UMLS SN integration.
- Consistency of SN Type combinations: As many UMLS concepts are categorized by more than one semantic type, there consistency against BioTop should be checked, based on the SN-BioTop map. To this end, all combinations of semantic types observed in Metathesaurus concepts must be identified and then attached to the ontology.

4 RESULTS

4.1 Mapping of UMLS semantic types

DL-based ontologies are hierarchies of types (classes) that can be instantiated by particular entities only. According to [McCray2002] we can consider the SN as a hierarchy of upper-level classes (regardless of the naming of some of the types that suggest a meta-level interpretation, e.g. the type *Functional Concept*). The categorization relation (that attaches UMLS concepts to SN types) can therefore be mostly interpreted as a taxonomic subsumption relation (*is-a*). Exceptions include geographical locations and a few other true instances, e.g., laws and persons. In these cases the categorization relation is to be interpreted as an *instance-of* relation.

The mapping was done as follows. First of all, the taxonomic tree of the UMLS SN types was recreated in an OWL file (SN.OWL) by expressing the hierarchical relations (*isa*) as OWL subclasses. No further assumptions were made. Especially, no partitions were introduced, as the source and its documentation do not make any statements as to whether semantic types are mutually exclusive. Based on the textual (SN, BioTop) and the formal (BioTop) definitions available we the attempted to map each Semantic Type to BioTop. Lexical mapping criteria were not used. In cases of doubt, domain experts were consulted. The mapping was performed in close collaboration between WikiProtein and BioTop curators. At several occasions, problems encountered when accommodating semantic types in BioTop were discussed in face to face meetings, conference calls, and e-mail discussions. In controversial cases other existing ontologies, e.g. . OBI, were consulted. For the mapping a new OWL bridging file was created that referenced both resources with owl:imports statements using the Protégé 4 ontology editor[7]. This allowed us to bring two resources together that were respectively out of our direct control and and to introduce new assertions linking them.

Mapping the semantic types of the Semantic Network to BioTop the following cases could be distinguished.

- Direct match: the ST is equivalent to a class in BioTop, or the difference is small enough that creating a separate new class alongside an existing one would not be justified; e.g., *Plant* in BioTop has the exact same meaning as in the SN.
- Restriction: no BioTop class is a straight match for the ST, but it can be defined by restricting an existing BioTop class, for example,
  AnatomicalAbnormality is mapped to:
  OrganismPart ⊓ ∃ **bearerOf**.*PathologicalCondition*,
  where *OrganismPart* and *PathologicalCondition* are existing BioTop classes and **bearerOf** is an existing BioTop relation;
- Union: if the ST can not be defined by a single class, but it can by the union of more than one..The contents of the union may be any combination of the previously described types. E.g., the SN type *Gene or Genome was* mapped to the expression biotop:Gene ⊔ biotop:Genome.
- Out of scope: the semantic type cannot be expressed using any of the options above; the immediate solution was to create a new class inside the mapping file itself, defined as the subclass of an existing BioTop class, and map the ST to this new class. In the incremental mapping / BioTop redesign process, all ST

[7] http://protege.stanford.edu/

leaf nodes (but two) introduced this way were recreated in BioTop. The non-matching semantic types (e.g. "daily or recreational activity") were mapped to a more general BioTop class.

- No match: the ST is regarded meaningless for BioTop in one of the following cases: its definition does not sufficiently differentiate it from its parent, it is too abstract, or it is only included in the SN as a "housekeeping" node in order to group more meaningful child nodes. For example, *Chemical Viewed Functionally* has a meta class meaning (it groups UMLS concepts, but is useless as a distinguishing criterion for their individuals) which cannot be represented by BioTop. Leaving the class undefined allows for the existing subsumption hierarchy of the SN to reason up to the nearest parent that does have a mapping, in this case *Substance*. Most semantic types on an upper level have nebulous definitions and do not coincide with any BioTop class, e.g. *Idea or concept* ("An abstract concept, such as a social, religious or philosophical concept."), the definition of which seems not plausible to its subtypes, e.g. *Geographic Area*.

The names, textual definitions and the hierarchical context of SN types created mapping difficulties in many cases. For instance, the ontologically crisp distinction between function and process is mixed up in the SN. So does the type *Phenomenon or Process* subsume *Pathologic Function*, which subsumes, e.g., *Neoplastic Process*. As a result we did not map a number of upper-level types. An example is Group, as it subsumes Family Group which, according to its definition, stands for a role, such as "parent" or "only child". Others were mapped to quite complex expressions with disjunctions and exclusions.

4.2 Interpretation and Mapping of UMLS semantic relations

The treatment of UMLS SN semantic relations turned out to be more complicated thus requiring a two-step approach: Firstly, they have to be semantically interpreted and properly built into an OWL-DL model, and secondly they will be mapped to BioTop. Their simple interpretation as description logics relations (object properties) is semantically problematic as SN relations range over semantic types (i.e. instantiable classes) whereas object properties range over individual entities. Such an interpretation of concept to concept relations in the light of formal logic has been repeatedly discussed in the recent years [Smith2005]. E.g., five different possible interpretations of SN triples are discussed in [Kashiap2003].

For most UMLS semantic relations there is a quite complex arrangement of domain and range restrictions, in which certain range restrictions are only valid with certain domain restrictions. For instance, the UMLS SN restricts the domain of the *treats* relation to drugs and physicians, and its range to patients and diseases (among others). However, it does not allow the combination of drug and patient, or health professional and disease[8].

Domain \ Range	*Drug*	*Physician*
Disease	allowed	disallowed
Person	disallowed	allowed

[8] For the sake of understandability the example is simplified and does not use the lengthy UMLS SN names.

We could, of course, ignore this and take simply the union of the extension of the UMLS concepts as the restriction of new BioTop relations that have to be included into the ontology. Thus we would have to accept unintended models, e.g. that a drug treats a person.

In our OWL-DL interpretation, however, we proceeded differently and introduced subrelations, in the following style (again simplified):

treatsMED ⊑ *treats* (domain: *Drug*, range: *Disease*)
treatsPHY ⊑ *treats* (domain: *Physician*, range: *Person*)

Doing so, we obtained a total number of 210 object properties. However, we have to acknowledge that this is a rather cosmetic solution, as this model is only able to reject unwanted assertions if the specialized relations are used but not if the general relations are used. Furthermore, by lack of disjointness statements in the class hierarchy it cannot even be rejected that, e.g. something is both a drug and a physician. This is, however, not a fault of the representation language but an underspecification of the UMLS SN.

A better solution would be the following, as it achieves the desired result without the creation of subrelations.

$Drug \sqsubseteq \forall \mathbf{treats}.Disease$
$Physician \sqsubseteq \forall \mathbf{treats}.Person$

Together with:

$\exists \mathbf{treats}.Disease \sqsubseteq Drug$
$\exists \mathbf{treats}.Person \sqsubseteq Physician$

However, this solution uses general concept inclusions (GCIs). Although part of the OWL DL specification they were not supported by our tools.

Any of the sketched models of representing semantic relations, however, faces a severe problem when it comes to the mapping to BioTop, as the latter includes only a relatively low number of relations. Enhancing BioTop by the whole array of SN relations would conflict with its design principle to keep the set of relations small and restrict them to those that are used in BioTop class definitions. This is not the case with most SN relations: *treats*, *interacts*, *diagnoses* etc. Instead, BioTop contains, in its *Processual Entity* branch classes such as *Treating*, *Interacting*, etc…, which convey the same meaning and can be regarded as reifications:

$TreatingPerson \sqsubseteq Action \sqcap$
$\quad \exists\ \mathbf{has\_agent}.\ Physician \sqcap \exists\ \mathbf{has\_patient}.\ Person \sqcap$
$\quad \forall\ \mathbf{has\_agent}.\ Physician \sqcap \forall\ \mathbf{has\_patient}.\ Person$
$TreatingDisease \sqsubseteq Action \sqcap$
$\quad \exists\ \mathbf{has\_agent}.\ Drug \sqcap \exists\ \mathbf{has\_\ patient}.\ Disease \sqcap$
$\quad \forall\ \mathbf{has\_agent}.\ Drug \sqcap \forall\ \mathbf{has\_\ patient}.\ Disease$
$Treating \equiv TreatingPerson \sqcup TreatingDisease$

We therefore decided to map – in an alternative approach – the SN relational constraints – expressed as triples – such as

D_1 REL R_1, D_2 REL R_2, D_3 REL R_3, …, D_n REL R_n

(D_i referring to domain and R_i to range) to an equally uncomplicated DL formula.

To this end we do not need to create new DL relations (which would contradict the DL design principles), but simplify the above formula:

$REL_1 \sqsubseteq \forall\ \mathbf{has\_domain}.\ D_1 \sqcap \forall\ \mathbf{has\_\ range}.\ R_1$
$REL_2 \sqsubseteq \forall\ \mathbf{has\_domain}.\ D_2 \sqcap \forall\ \mathbf{has\_\ range}.\ R_2$
$REL_3 \sqsubseteq \forall\ \mathbf{has\_domain}.\ D_3 \sqcap \forall\ \mathbf{has\_\ range}.\ R_3$
…
$REL_n \sqsubseteq \forall\ \mathbf{has\_domain}.\ D_n \sqcap \forall\ \mathbf{has\_\ range}.\ R_n$

$REL \equiv REL_1 \sqcup REL_2 \sqcup REL_3 \sqcup \ldots \sqcup REL_n$

has_domain and **has_ range** are then mapped to biotop: **has_agent** and biotop:**has_patient.**

Of course, the agent / patient reading does not make sense with spatial or temporal relations.

Finally, there are SN relations that cannot be expressed as relations between particulars because they simply do not relate anything at the level of particulars. The prototypical example is "prevent", such as in the statement "contraceptive drugs prevent pregnancy". On a UMLS concept level it is, without doubt, sensible to express this in a relational form, such as

"*prevents (contraceptive drugs; pregnancy)*"

Such a close-to-human-language assertion on prevention carries several implicit assumptions that must be made clear before expressing it via an ontology: Preventing pregnancy does not exclude the possibility of becoming pregnant but it brings about a strong risk reduction. Furthermore, there is a both a temporal and dose association between the drug and the risk. We can therefore rephrase "Contraceptive drugs prevent pregnancy" as follows: "The administration of contraceptive drugs of an adequate dose and regularity to a woman reduces her pregnancy risk within a defined timeframe", or more simply: "The administration of contraceptive drugs to a woman reduces her pregnancy risk within a defined timeframe". We could express this as follows::

PregnancyRiskReductionBySubstanceIntake $\sqsubseteq$
Action $\sqcap \exists$ **has_agent**.*Substance* $\sqcap$
$\forall$**has_agent**.*Substance* $\sqcap$
$\exists$ **has_patient**. (*Risk* $\sqcap$ ($\exists$ **inheres_in**. *Organism* $\sqcap$
$\forall$ **inheres_in**. *Organism***)**
$\sqcap$ $\forall$**risk_of**.*Pregnancy)*

This digression illustrates the difficulty if not impossibility of an ontologically precise formal reconstruction of seemingly simple close-to-language predicates.

For the semantic relationship mapping we proceeded the following way: All relationships were reified (i.e. expressed as classes) and added as OWL classes using value restrictions on the roles **has_agent** and **has_patient**. Those relationships which had a direct correlate in BioTop (i.e., the SN spatiotemporal relationships) were additionally mapped directly to BioTop relationships (object properties). In both cases the domain and range-specific subrelations were accounted for by additional subclasses / subrelations (in analogy to the "Treating" example above). The reification classes were furthermore provided with so-called covering axioms that assure the enforcement of one of the child classes with their restrictions. Again, no mappings were performed for some upper-level relationships (and, accordingly to upper-level reification classes), for the same reasons as explained for the type hierarchy. The final result of the mapping of each Semantic Type to BioTop yielded 70 equivalence statements in the mapping ontology, the OWL reconstruction of the UMLS SN with about 350 classes and 550 axioms as well as the upgrade of BioTop from 200 to about 300 classes, 30 to 50 object properties, and from 470 to roughly 750 axioms.

4.3 Assessment results

The mapping exercise constituted an ideal testbed for the ongoing quality assurance and formative evaluation of BioTop. Due to the constant need of inconsistency checking and resolving, many hidden errors in BioTop were detected, especially faulty disjointness axioms (e.g. *Organic Chemical* was disjoint from *Carbohydrate*), unrecognized ambiguities (e.g. *Sequence* as information entity vs. sequence as molecular structure), as well as granularity mismatches (e.g. *Chromosome* as molecule). This highly beneficial maintenance work was however very time consuming, totaling at least one person year, divided among five modelers. A significant advance was achieved by the use of a new Protégé add-in that presents precise explanations of entailments in OWL ontologies [Horridge2008]. The justifications of classification errors visualized by this tool offered a great advantage in the process of inconsistency check and resolution.

The results of the named entity experiment are summarized in Table 1. Due to many ambiguities, the curator decided in only half of the cases to assign a clear judgment in the sense of semantically related vs. semantically unrelated. The comparison between the tables clearly demonstrates the dilemma. The UMLS Semantic Network shows a certain correlation with the curator's judgment but still produces many false negatives and false positives, BioTop – via the SN and the mapping ontology – rejects extremely few associations.

Table 1. Named entity co-occurrence results

	Expert judgment: related	Expert judgment: unrelated
SN: related	31	22
SN: unrelated	21	71
BioTop: related	50	89
BioTop: unrelated	2	4

The BioTop result shows the problem of the so-called open world semantics [Baader2007], i.e. all models are accepted unless they are explicitly falsified. In the case a description logics ontology is used for this kind of consistency check, the modelers have to be very meticulous in "filling the holes". On the other hand, it must be acknowledged that the OWL reconstruction of the idiosyncratic categorization in SN required many disjunctive statements which conveyed a relaxation of the domain and value restrictions. In any way, it is known to be difficult to keep an OWL model "water-proof" in this aspect, and OWL has recently been criticized that is generally ill-suited for tasks like schema validation [Rajsky2008]. However, we argue that this is less an inherent but rather a tooling problem, at least for description logics that support some kind of negation. Recent developments that facilitate the spotting of inconsistencies [Horridge2008] provide room for optimism. For BioTop, these results tell us that a thorough fault analysis is mandatory.

The mapping ontology produced better results in the experiment on multiple semantic type categorization, where we checked each occurring STY combination against the mapping ontology. In the 2008 Metathesaurus (totaling more than 1.80 Million concepts) release we found 397 different combinations types of two to four semantic types, occurring in about 220,000 UMLS concepts. The DL classifier recognized 89 combinations as inconsistent, affecting a total of 2,954 UMLS concepts. The most frequently occurring type combination was *Manufactured Object* with *Health Care Related Organization* (e.g. *Hospital* as building vs. organization).

5 RELATED WORK

There are many evidences in literature for converting thesauri, frame knowledge bases and ontologies from various representational formats into description logics. Examples are [Pisanelli1998] and [Schulz2001] for the UMLS, [Beck2003], [Dameron2005], and [Noy2008] for the Foundational Model of Anatomy, [Wroe2003] and [Egana2008] for the Gene Ontology, and [Heja2007] for ICD-10. What most of these approaches have in common is (i) that the mapping is not straightforward, (ii) it relies on several ontological basic assumptions that are not explicitly stated in the sources, e.g. on disjointness axioms, on the intended meaning and the algebraic properties of relationships, and (iii) that not all knowledge conveyed by the sources is expressible in description logics, due to the language constraints.

The UMLS SN was targeted by [Kashyap2003] who concluded that the logical interpretation of the semantic relations in the SN should depend on the application in which the ontology is to be used. More specifically, ontological aspects of the UMLS SN were discussed by [Schulze2004]. The latter authors acknowledge the importance of the SN for the semantic integration of terminology but spot a number of weaknesses future revisions should address. A major point of criticism is the mixture of concrete with abstract entities, real entities with "bauplan" entities, objects with their roles, functions and processes. This mainly coincides with our mapping experiences as described in 4.1. and 4.2.

6 CONCLUSION

We have described the ongoing development and improvement of a semantic resource, the life science ontology BioTop in the light of the mapping to the legacy UMLS semantic network. The purpose of this effort is to bring together the large amount of data categorized by the latter with the formal foundation of the former, using emerging standards and tools developed by the Semantic Web community. Semantic and terminological support is especially important for facilitating an opening of the curation process towards a broader community, which is the goal of WikiProteins, an ambitious annotation project in the context of which the present work was performed.

The alignment process of a formal ontology with a relatively informal system of hierarchically ordered categories like the UMLS semantic network challenges the ontology engineer to formally reinterpret the latter and to overcome its ontological shortcomings. The logical machinery of description logics, implemented in reasoning engines, was an indispensable part of the mapping process, which, at the end of the day, not only provided a consistent mapping ontology but contributed, by large, to error detection and improvement of BioTop.

We described two assessment experiments. One or them, aiming at consistency checking of SN type combinations yielded good results that revealed hidden ambiguities of UMLS concepts. The other, however, yielded a poor result. It aimed at using the ontology for determining which UMLS concept pairs were closely related to each other. As a result, the mapping ontology rejected very few models, thus supporting the recent critique on the suitability of OWL for schema verification. It was disappointing because the modelers had spent a great effort in partitioning the BioTop ontology in order to antagonize the unwarranted effects of the open world assumption. This is an issue where more sophisticated tool support for OWL ontology construction and validation is desperately needed, in order to grant formal ontologies and logic-based reasoning a central place in future high-throughput and high-impact life sciences knowledge management technologies.

ACKNOWLEDGEMENTS

This work was supported by the EC STREP project "BOOTStrep" (FP6 – 028099) and by the Intramural Research Program of the National Institutes of Health (NIH), National Library of Medicine (NLM). We also thank Martin Boeker (Freiburg), Holger Stenzhorn (Freiburg) for their BioTop maintenance efforts.

REFERENCES

[Ashburner2000] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock (2000). Gene Ontology: tool for the unification of biology. Nature Genetics 25:25–29

[Baader2007] F. Baader, D. Calvanese, D.L. McGuinness, D. Nardi, and P.F. Patel–Schneider, eds. (2007). The Description Logic Handbook. 2nd ed. Theory, Implementation, and Applications, Cambridge, U.K.: Cambridge University Press.

[Beck2003] R. Beck and S. Schulz. Logic-based remodeling of the DIGITAL ANATOMIST Foundational Model. In AMIA 2003–Proceedings of the Annual Symposium of the American Medical Informatics Association, 687–691. Washington, D.C., November 8-12, 2003. Philadelphia, PA: Hanley & Belfus, 2003.

[BernersLee2001] T. Berners-Lee, J. A. Hendler, Ora Lassila, The Semantic Web, Scientific American, 284(5):34–43, May 2001.

[Bodenreider2004] O. Bodenreider. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267–70.

[Dameron2005] O. Dameron, D.L. Rubin, and M.A. Musen (2005). Challenges in Converting Frame-Based Ontology into OWL: the Foundational Model of Anatomy Case-Study. In AMIA 2005 – Proceedings of the Annual Symposium of the American Medical Informatics Association, 181–185

[DeKeizer2000a] N. F. de Keizer, A. Abu-Hanna, and J. H. Zwetsloot-Schonk. Understanding terminological systems. i: Terminology and typology. Methods of Information in Medicine, 39(1):16–21, 2000.

[DeKeizer2000b] N. F. de Keizer and A. Abu-Hanna. Understanding terminological systems. ii: Experience with conceptual and formal representation of structure. Methods of Information in Medicine, 39(1):22–29, 2000.

[Egana2008] M. Egaña Aranguren, C. Wroe, C. Goble, and R. Stevens (2008). In situ migration of handcrafted ontologies to reason-able forms. *Data Knowl. Eng.* 66, 1 (Jul. 2008), 147–162.

[Gangemi2002] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider, L. (2002). Sweetening ontologies with Dolce. In: A. Gómez-Pérez, R.V.Benjamins (eds.) . Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web. Proceedings of the 13th International Conference – EKAW 2002; Berlin: Springer; 166 –181.

[Guarino1998] N. Guarino (1998). Formal Ontology in Information Systems. Proceedings of FOIS'98, Trento, Italy, 6–8 June 1998. Amsterdam, IOS Press, 3–15.

[Haarslev2003] V. Haarslev and R. Möller. Racer: An OWL reasoning agent for the semantic web (2003) Proc. of the International Workshop on Applications, Products and Services of Web-based Support Systems, in conjunction with 2003 IEEE/WIC International Conference on Web Intelligence.

[Heja2007] G. Héja, G. Surján, G. Lukácsy, P.Pallinger, M. Gergely M. GALEN based formal representation of ICD10. International Journal of Medical Informatics 2007 Feb-Mar;76(2–3):118–23.

[Heller2004] B. Heller and H. Herre (2004). Ontological Categories in GOL. Axiomathes ,14, 57–76. Kluwer Academic Publishers.

[Hoehndorf2009] R. Hoehndorf. GFO-Bio: A biomedical core ontology. http://onto.eva.mpg.de/gfo-bio.html, last accessed: Jan 4, 2009.

[Horridge2008] M. Horridge, B. Parsia and U. Sattler. . "Laconic and Precise Justifications in OWL." International Semantic Web Conference 2008, Karlsruhe, Germany.

[Horrocks2003] I. Horrocks, P. F. Patel-Schneider and F. van Harmelen. From SHIQ and RDF to OWL: The making of a Web ontology language. Journal of Web Semantics, 1(1):7–26, 2003.

[ISO2000] International Organization for Standardization (ISO) (2000): ISO 1087 – 1: Terminology work – Vocabulary – Part 1: Theory and applications, Geneva, Switzerland.

[Jelier2005] R. Jelier, G. Jenster, L.C.J. Dorssers, C.C. Van der Eijk, E.M. van Mulligen, B. Mons, and J.A. Kors (2005) Co-occurrence based meta-analysis of scientific texts: retrieving biological relationships between genes. Bioinformatics 21:9, 2049–2058

[Kashyap2003] V. Kashyap and A. Borgida. Representing the UMLS Semantic Network using OWL: (Or "What's in a Semantic Web link?"). In: Fensel, D., Sycara, K., Mylopoulos, J. (Eds.), The SemanticWeb–ISWC, 2003, Springer-Verlag, Heidelberg. 1–16.

[Kusnierczyk2006] W. Kuśnierczyk (2006). Nontological Engineering. Formal Ontology in Information Systems. Proceedings of the 4th International Conference FOIS 2006, 39–50.

[Kumar2004] A. Kumar, B. Smith, S. Schulze-Kremer: Revising the UMLS Semantic Network", in: Medinfo 2004, San Francisco, 2004

[Masolo2003] C. Masolo, S. Borgo, A. Gangemi, N. Guarino, A. Oltramari (2003). WonderWeb Deliverable D18 Ontology Library. Infrastructure for the Semantic Web. http: //wonderweb.semanticweb.org, last accessed Jan 11th, 2008.

[McCray1995] A.T. McCray and S. J. Nelson. The representation of meaning in the UMLS. Methods of Information in Medicine, 34(1/2):193–201, 1995.

[McCray2002] A.T. McCray and O. Bodenreider. A conceptual framework for the biomedical domain In: R. Green, C.A. Bean, and S.H. Myaeng SH, editors. The semantics of relationships: an interdisciplinary perspective. Dordrecht; Boston: Kluwer Academic Publishers; 2002, 181–198.

[McCray2003] A.T. McCray. An upper-level ontology for the biomedical domain. Comparative and Functional Genomics. 2003;4(1):80–84.

[Mulder2008] N.J. Mulder , P. Kersey, M. Pruess, R. Apweiler. In silico characterization of proteins: UniProt, InterPro and Integr8. Molecular Biotechnology 2008 Feb;38(2):165–177.

[Mons2008] B. Mons, M. Ashburner, C. Chichester, E. van Mulligen, M. Weeber, J. den Dunnen, G.J. van Ommen, M. Musen, M. Cockerill, H. Hermjakob, A. Mons , A. Packer, R. Pacheco, S. Lewis, A. Berkeley, W. Melton, N. Barris, J. Wales, G. Meijssen, E. Moeller, P. J. Roes, K. Borner, and A. Bairoch. Calling on a million minds for community annotation in WikiProteins. Genome Biology 2008 May 28, 9(5):R89

[Noy2008] N.F. Noy, D.L. Rubin. Translating the Foundational Model of Anatomy into OWL. Journal of Web Semantics 2008;6(2):133–136.

[Park2006] J. C. Park and J.-J. Kim. Named entity recognition. In S. Ananiadou and J. McNaught, editors, Text Mining for Biology and Biomedicine, 121–142. Artech House, Boston, 2006.

[Pease2008] A. Pease. The Suggested Upper Merged Ontology (SUMO) 2008. http://www.ontologyportal.org/

[Pisanelli1998] D.M. Pisanelli, A. Gangemi, and G. Steve, An ontological analysis of the UMLS Metathesaurus. In AMIA'98 – Proceedings of the 1998 AMIA Annual Fall Symposium. A Paradigm Shift in Health Care Information Systems: Clinical Infrastructures for the 21st Century, Orlando, FL, November 7–11, 1998, Hanley and Belfus, Philadelphia, PA, 1998, pp. 810–814.

[Quine1948] O. Quine (1948). On What There Is. Review of Metaphysics.

[Rajsky2008] P. Rajsky. Canonical data model - Using industry standard data models. Theory and Practice of System Integration. Weblog at http://it.toolbox.com/blogs/system-integration-theory

[Rector2006] A. Rector, R. Stevens and J. Rogers (2006). Simple Bio Upper Ontology. [http: //www.cs.man.ac.uk/~rector/ontologies/simple-top-bio] – Last accessed: Jan 6 2009.

[Ruttenberg2007] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, O. Bodenreider, H. Chen, D. Doherty, K. Forsberg, Y. Gao, V. Kashyap, J. Kinoshita, J. Luciano, M. Scott Marshall, C. Ogbuji, J. Rees, S. Stephens, G. T. Wong, E. Wu, D. Zaccagnini, T. Hongsermeier, E. Neumann, I. Herman, K.-H. Cheung. Advancing translational research with the Semantic Web. BMC Bioinformatics. 2007 May 9;8 Suppl 3:S2.

[Sagotsky2008] J. A. Sagotsky, L. Zhang, Z. Wang, S. Martin, T. S. Deisboeck. Life Sciences and the web: a new era for collaboration. Molecular Systems Biology 2008; 4: 201.

[Schuemie2007] M. Schuemie, R. Jelier, and J.A. Kors. Peregrine: Lightweight gene name normalization by dictionary lookup. Proceedings of the Biocreative 2 workshop 2007 April 23–25, Madrid, 131–140

[Schulz2001] S. Schulz and U. Hahn. Medical knowledge reengineering--converting major portions of the UMLS into a terminological knowledge base. Int J Med Inform. 2001 Dec;64(2–3):207–21.

[Schulz2006] S. Schulz, E. Beisswanger, U. Hahn, J. Wermter, H. Stenzhorn, and A. Kumar (2006). From GENIA to BioTop – Towards a Top-level Ontology for Biology. 4th International Conference on Formal Ontology in Information Systems (FOIS 2006), Baltimore, USA, November 2006, 103–114.

[Schulze2004] S. Schulze-Kremer, B. Smith and A. Kumar. Revising the UMLS Semantic Network. http://ontology.buffalo.edu/medo/UMLS_SN.pdf, 2004

[Sirin2007] E. Sirin, B. Parsia, B. Cuenca Grau, A. Kalyanpur, Y. Katz. Pellet: A practical OWL-DL reasoner. Web Semantics: Science, Services and Agents on the World Wide Web, Volume 5, Issue 2, Software Engineering and the Semantic Web, June 2007, 51–53, ISSN 1570–8268.

[Smith2005] B. Smith, W. Ceusters, B. Klagges, J. Köhler, A. Kumar, J. Lomax, C. Mungall, F. Neuhaus, A. L. Rector, C. Rosse (2005). Relations in Biomedical Ontologies, Genome Biology. 2005; 6 (5).

[Smith2007] B. Smith, M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, OBI Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.A. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel , and S. Lewis (2007). The OBO Foundry: Coordinated Evolution of Ontologies to Support Biomedical Data Integration. Nature Biotechnology 25 (11): 1251–1255.

[Smith2007a] B. Barry and P. Grenon. Basic Formal Ontology. http://www.ifomis.uni-saarland.de/bfo/,

[SNOMED2009] SNOMED Clinical Terms. Copenhagen: International Health Standards Development Organisation, 2009. http://www.ihtsdo.org

[Stevens2000] R. Stevens, C.A. Goble, and S. Bechhofer (2000) Ontology-based knowledge representation for bioinformatics. Briefings in Bioinformatics 1(4):398–416

[Tsarkov2006] D. Tsarkov and I. Horrocks. FaCT++ Description Logic Reasoner: System Description. Lecture Notes in Computer Science, Vol. 4130/2006. 2006, 292–297.

[UMLS2009] UMLS. *Unified Medical Language System*. Bethesda, MD: National Library of Medicine, 2009.

[Wroe2003] C.J. Wroe, , R. Stevens, C.A. Goble, M. Ashburner (2003) . A Methodology to Migrate the Gene Ontology to a Description Logic Environment Using DAML+OIL. Pacific Symposium on Biocomputing 8:624–635